

# Building Infrastructure for Data-Driven Research

Dr. Philipp Zumstein  
Mannheim University Library



2017-03-15 Social Science Data Lab, Mannheim

[https://github.com/SocialScienceDataLab/  
building-infrastructure-for-data-driven-research](https://github.com/SocialScienceDataLab/building-infrastructure-for-data-driven-research)

Slides are Open Access, reuse them as



(this does not cover necessarily all the pictures;  
see individual attributions )

# Overview

- Data-driven Research
- Building Infrastructure
- OCR Workflow
- OCR Software
- Applications

# Data-driven Research

collected  
data

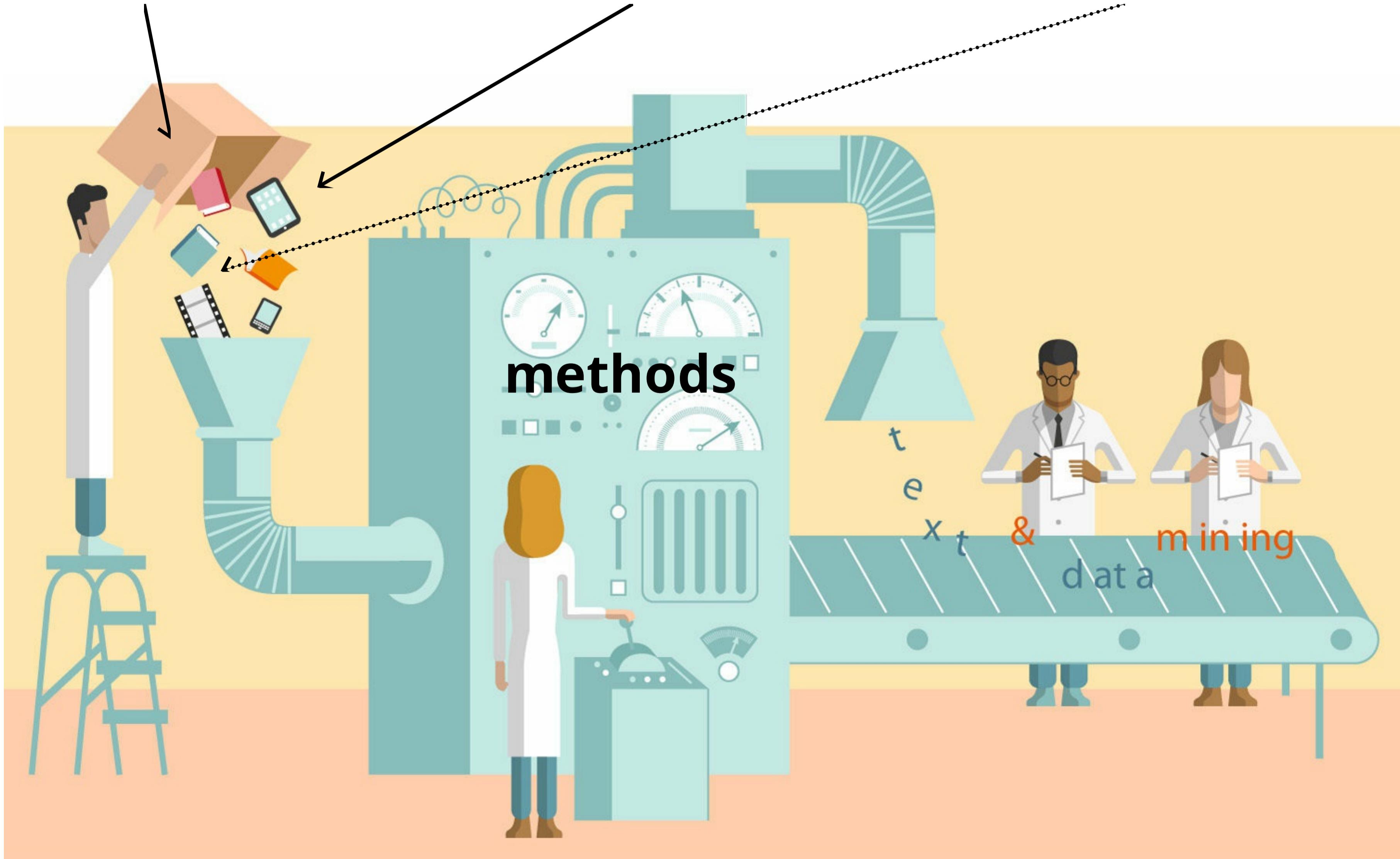
time

online  
data

§§

other  
data

?

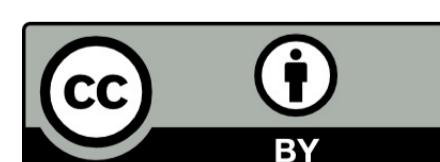
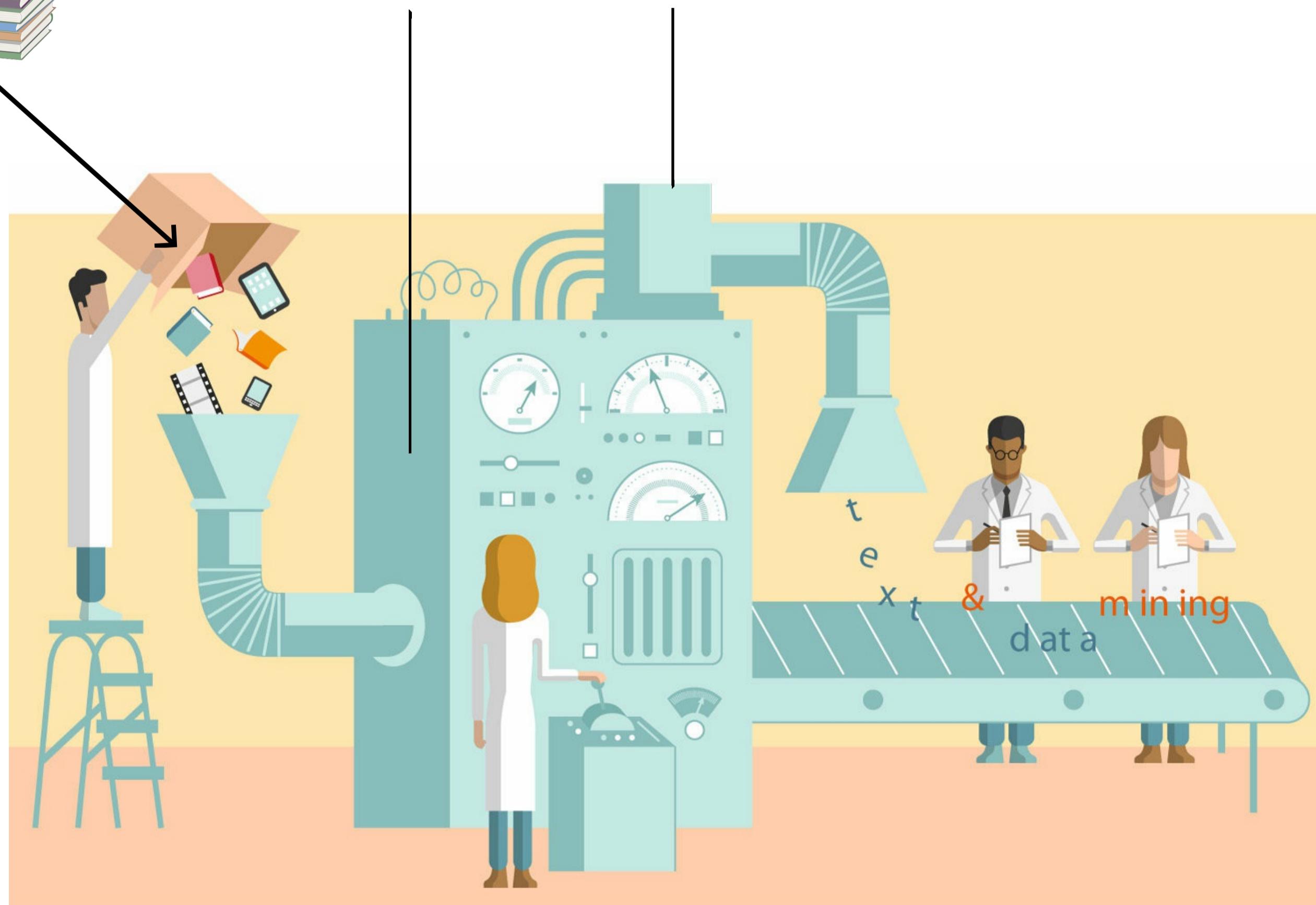


Copyright User (2013-06): Text and Data Mining (Original Illustration by Davide Bonazzi)  
<http://copyrightuser.org/topics/text-and-data-mining/>

What do you do  
with images containing  
text or printed  
books/newspapers as  
input?



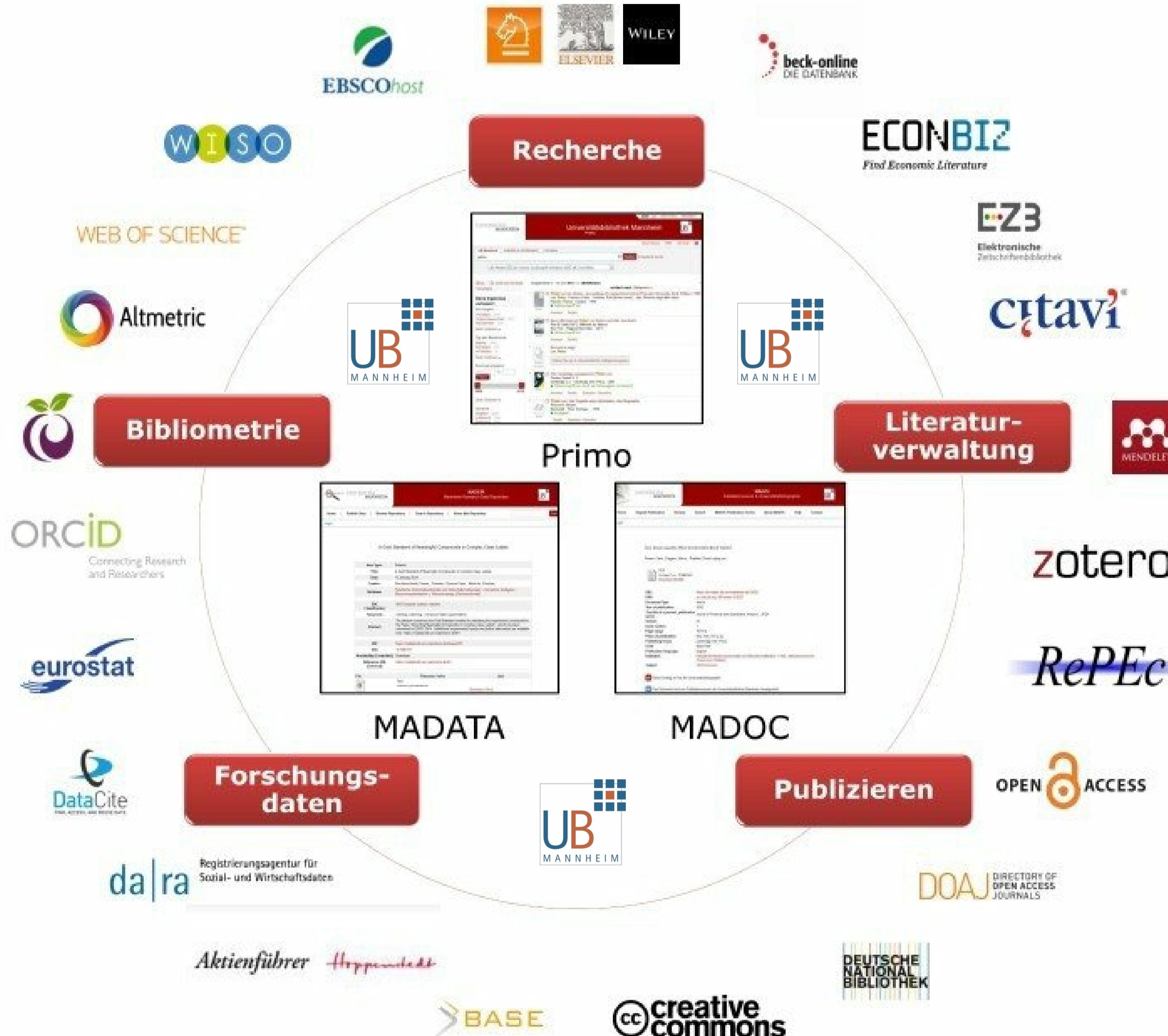
## Digitization, OCR, Structuring (infrastructure for research)



Copyright User (2013-06): Text and Data Mining (Original Illustration by Davide Bonazzi)  
<http://copyrightuser.org/topics/text-and-data-mining/>

# **Building Infrastructure**

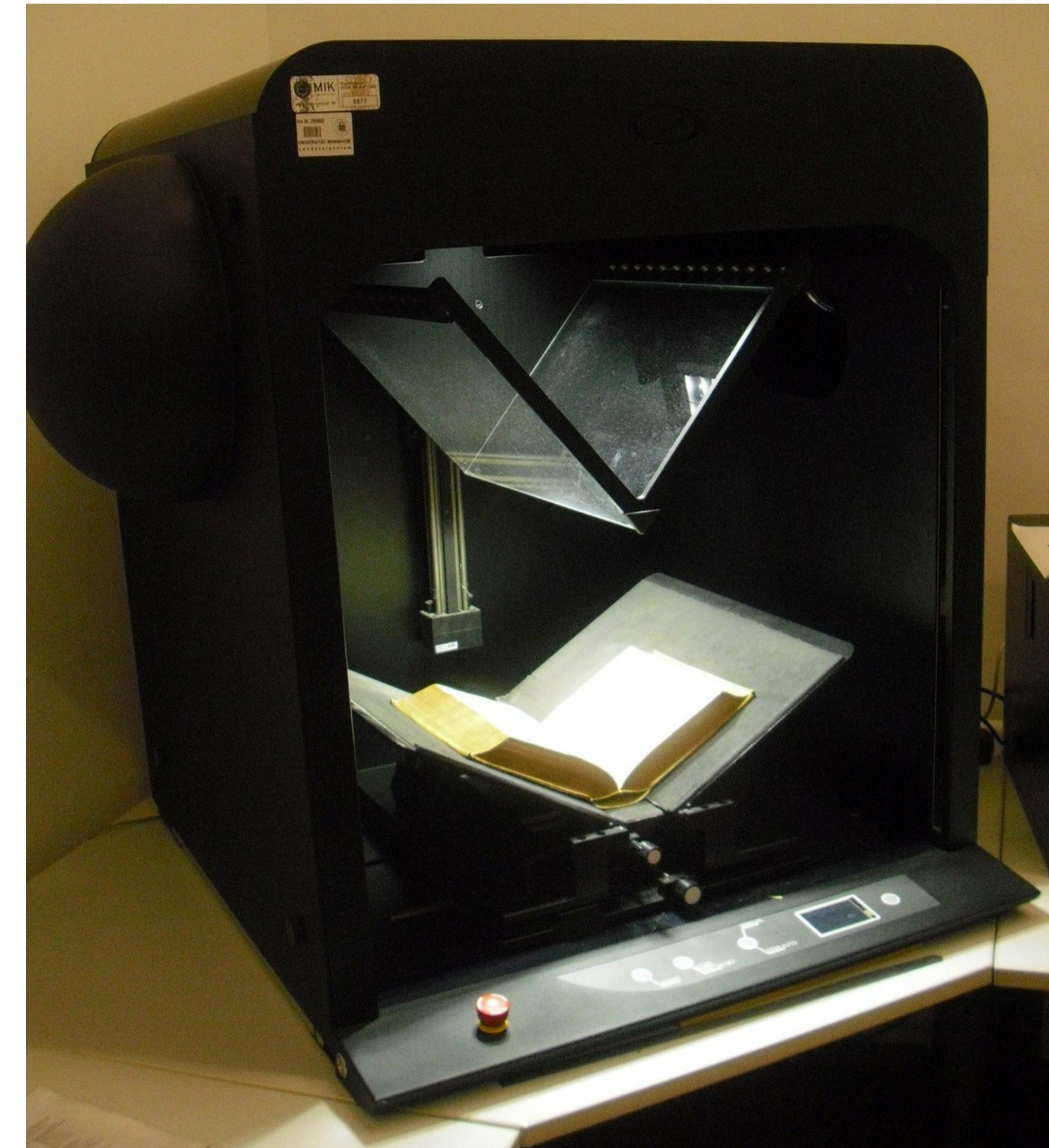
# Science Support from Library



# Infrastructure for Scanning



A1-Scanner for newspapers etc.



V-Scanner for rare, old, fragile books

# Digitization, "Data-ization"

## Our Digitization Infrastructure:

- V-scanner
- A1-scanner
- A2-scanner
- A3-scanner
- conservation checks and fixes

## Our Expertise:

- scanning workflow
- (manual) double-key-methods
- automatic text recognition (OCR)
- digitizing microfiche, microfilm
- extracting information from CDs to a database
- structuring information
- metadata formats

# Infrastructure Projects

**Ancien Droit:** digitizing 800 books from the 17th/18th century from the collection of Desbillon with focus on the history of the "Ancien Droit" <https://digi.bib.uni-mannheim.de/>

**Aktienführer I+II:** digitizing the annualy published books "Aktienführer", extracting the data in a data base  
<https://digi.bib.uni-mannheim.de/aktienfuehrer/>

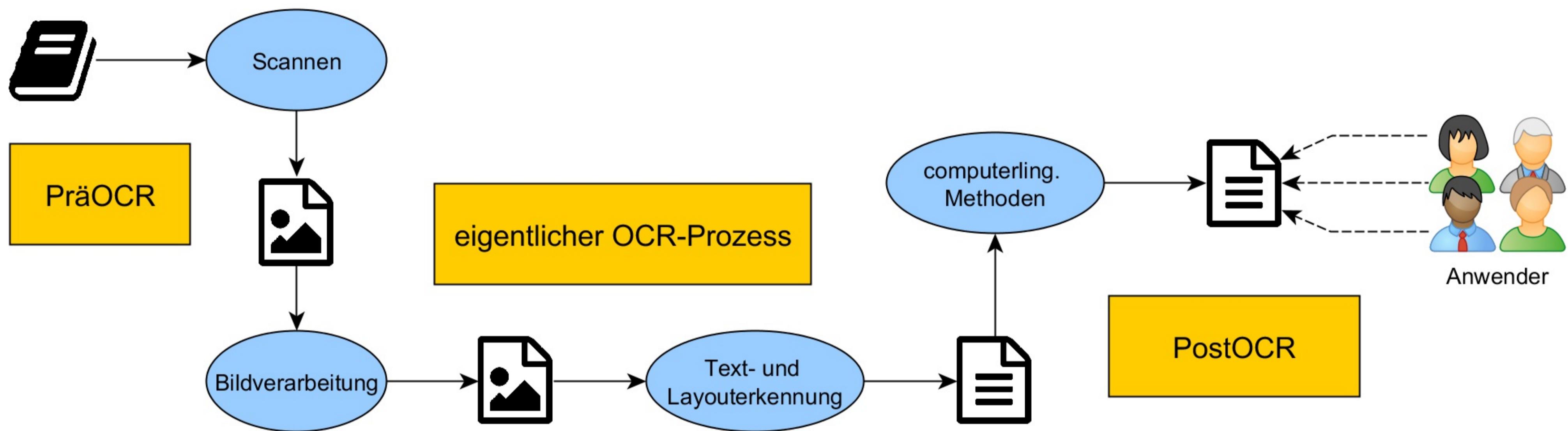
**Reichsanzeiger:** German newspaper (government gazette) from 1819 to 1945 <https://digi.bib.uni-mannheim.de/periodika/reichsanzeiger/>

**LOC-DB:** open, distributed infrastructure for cataloguing of citations <https://locdb.bib.uni-mannheim.de/>

**Infolis I+II:** connect research data and publications, text mining scientific articles, integration into different retrieval systems <http://infolis.github.io/>

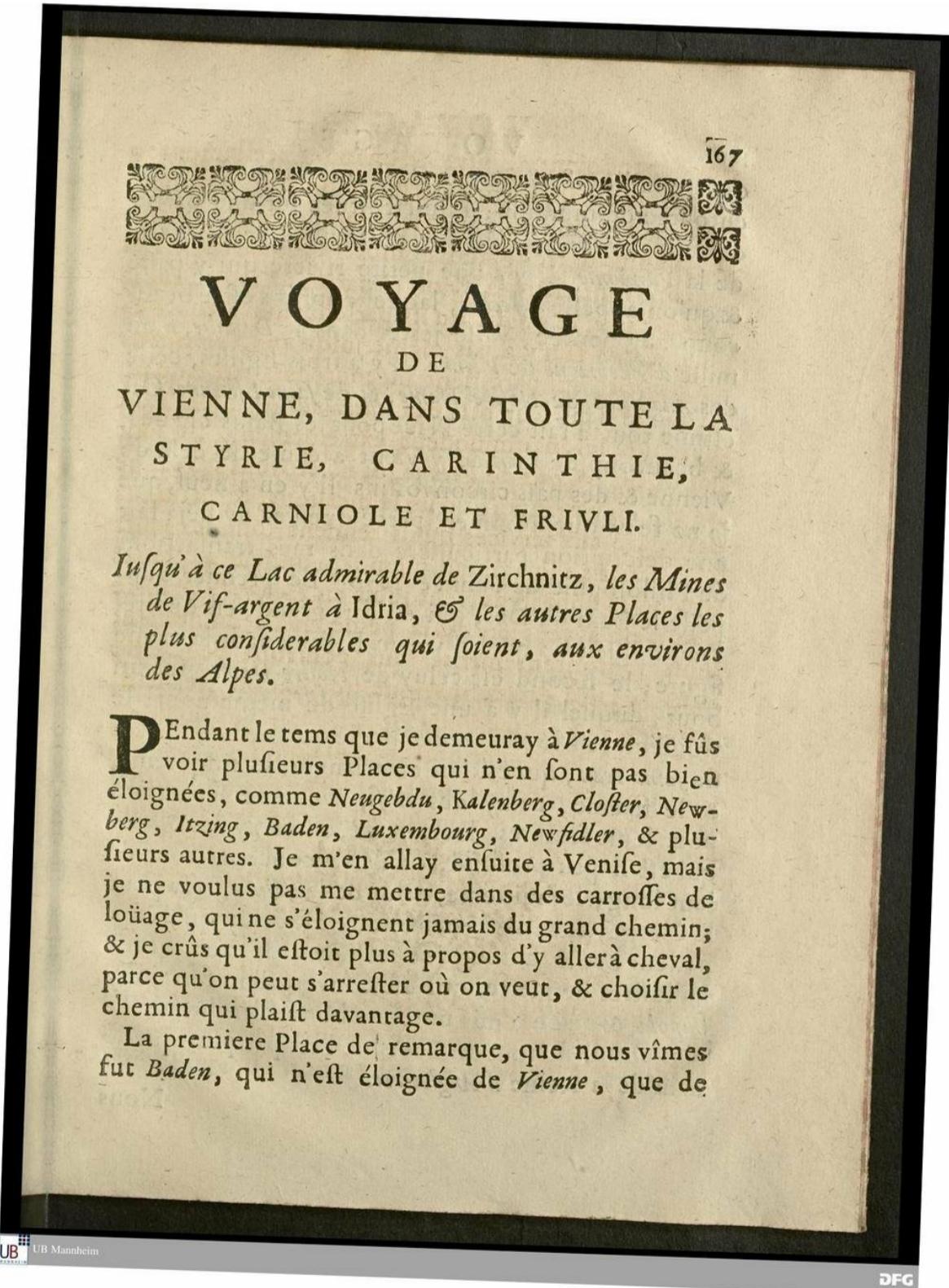
# OCR Workflow

# Workflow of OCR-Process

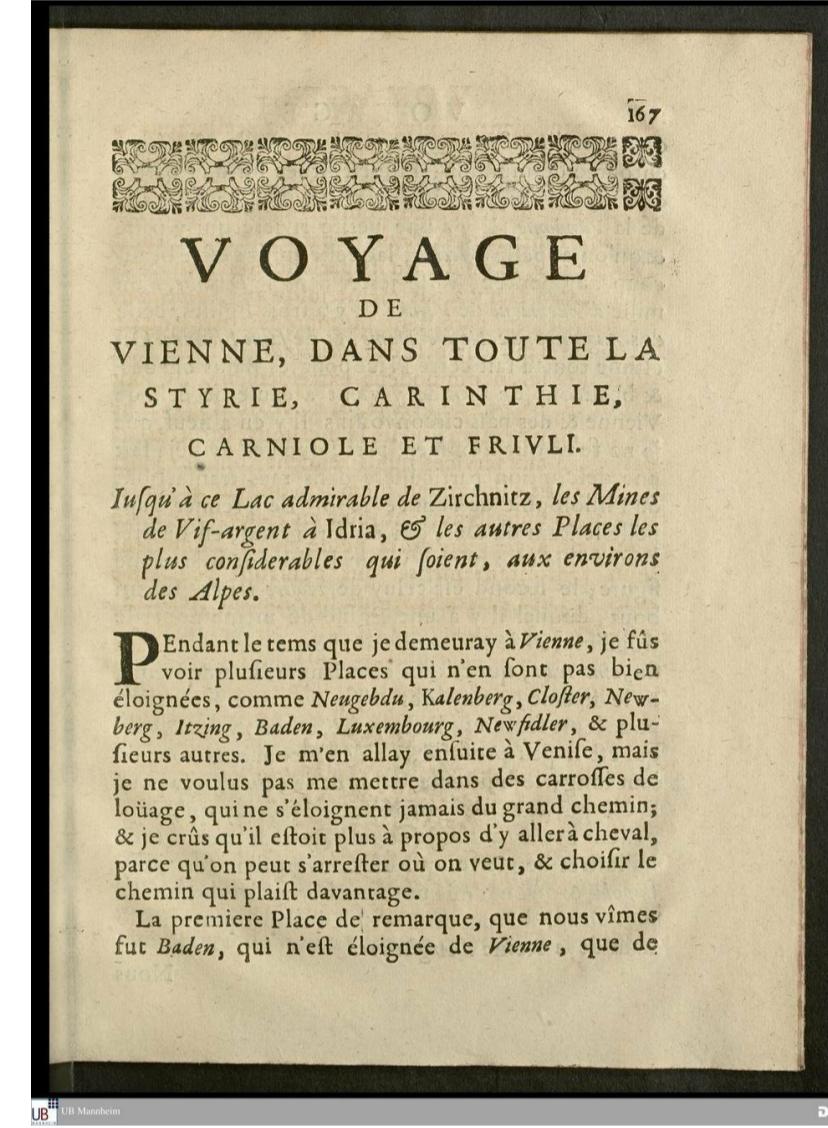


Baierer, Konstantin; Zumstein, Philipp (2016). Verbesserung der OCR in digitalen Sammlungen von Bibliotheken. *027.7 Zeitschrift für Bibliothekskultur / Journal for Library Culture*, v. 4, n. 2, p. 72-83.  
<https://doi.org/10.12685/027.7-4-2-155>

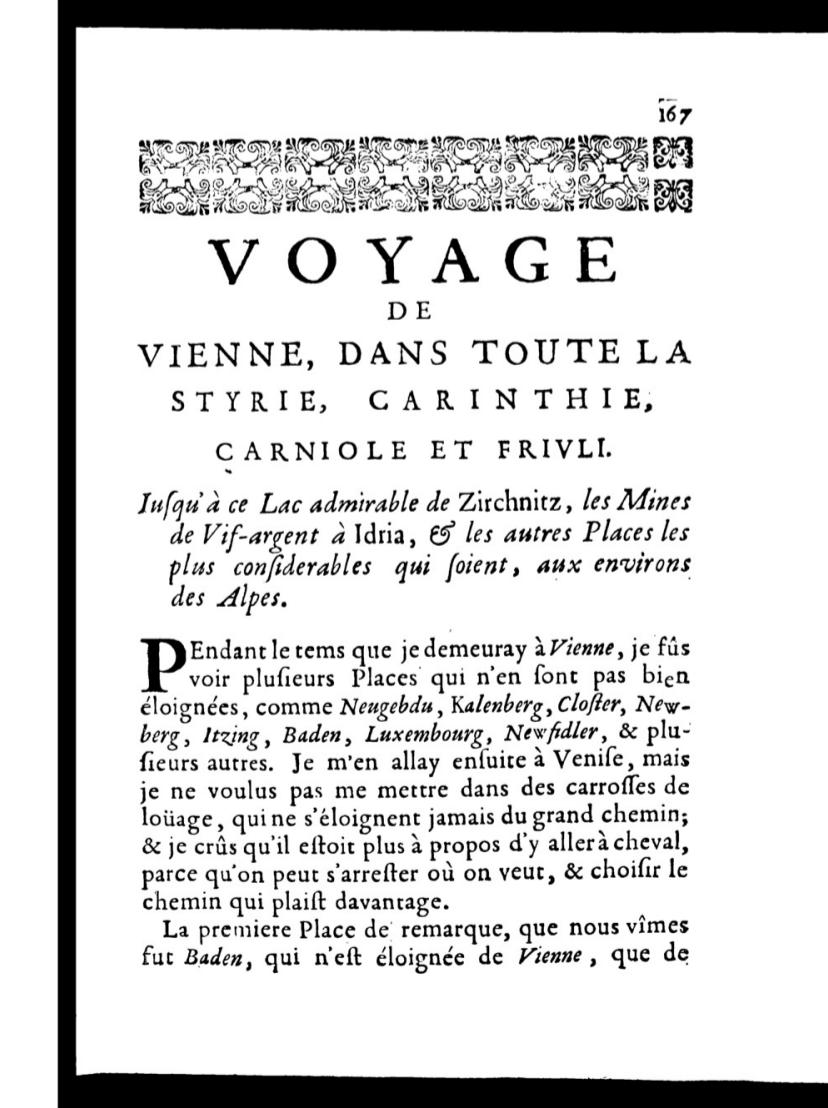
# Image Processing



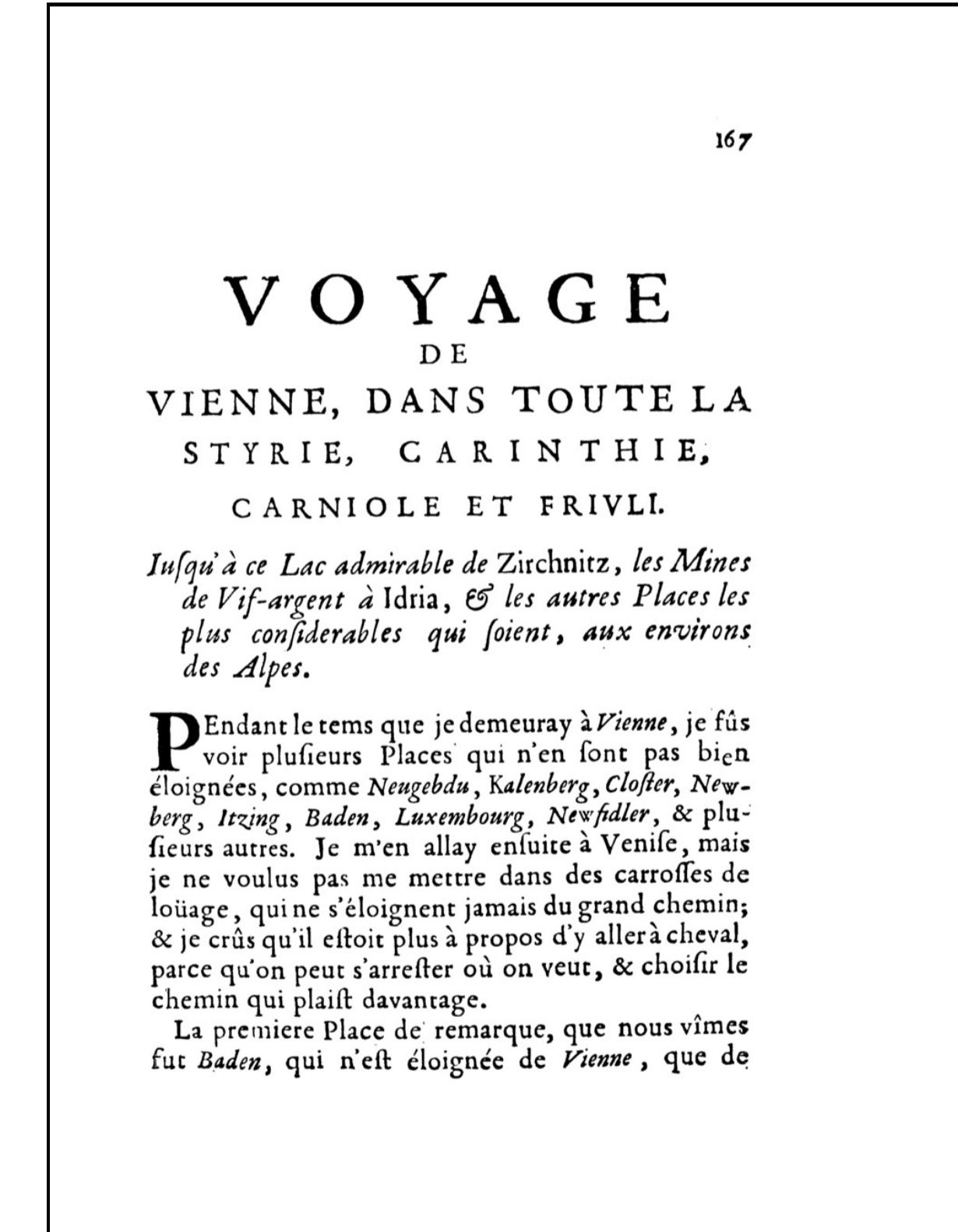
deskew  
dewarp



binarize



denoise  
despeckling



167  
VOYAGE  
DE  
VIENNE, DANS TOUTE LA  
STYRIE, CARENTHIE,  
CARNIOLE ET FRIVLI.

*Jusqu'à ce Lac admirable de Zirchnitz, les Mines de Vif-argent à Idria, & les autres Places les plus considérables qui soient, aux environs des Alpes.*

**P**endant le tems que je demeuray à Vienne, je fûs voir plusieurs Places qui n'en sont pas bien éloignées, comme Neugebdu, Kalenberg, Closter, Newberg, Itzing, Baden, Luxembourg, Newfdler, & plusieurs autres. Je m'en allay ensuite à Venise, mais je ne voulus pas me mettre dans des carrosses de louage, qui ne s'éloignent jamais du grand chemin; & je crus qu'il estoit plus à propos d'y aller à cheval, parce qu'on peut s'arrêter où on veut, & choisir le chemin qui plaist davantage.

La première Place de remarque, que nous vîmes fut Baden, qui n'est éloignée de Vienne, que de

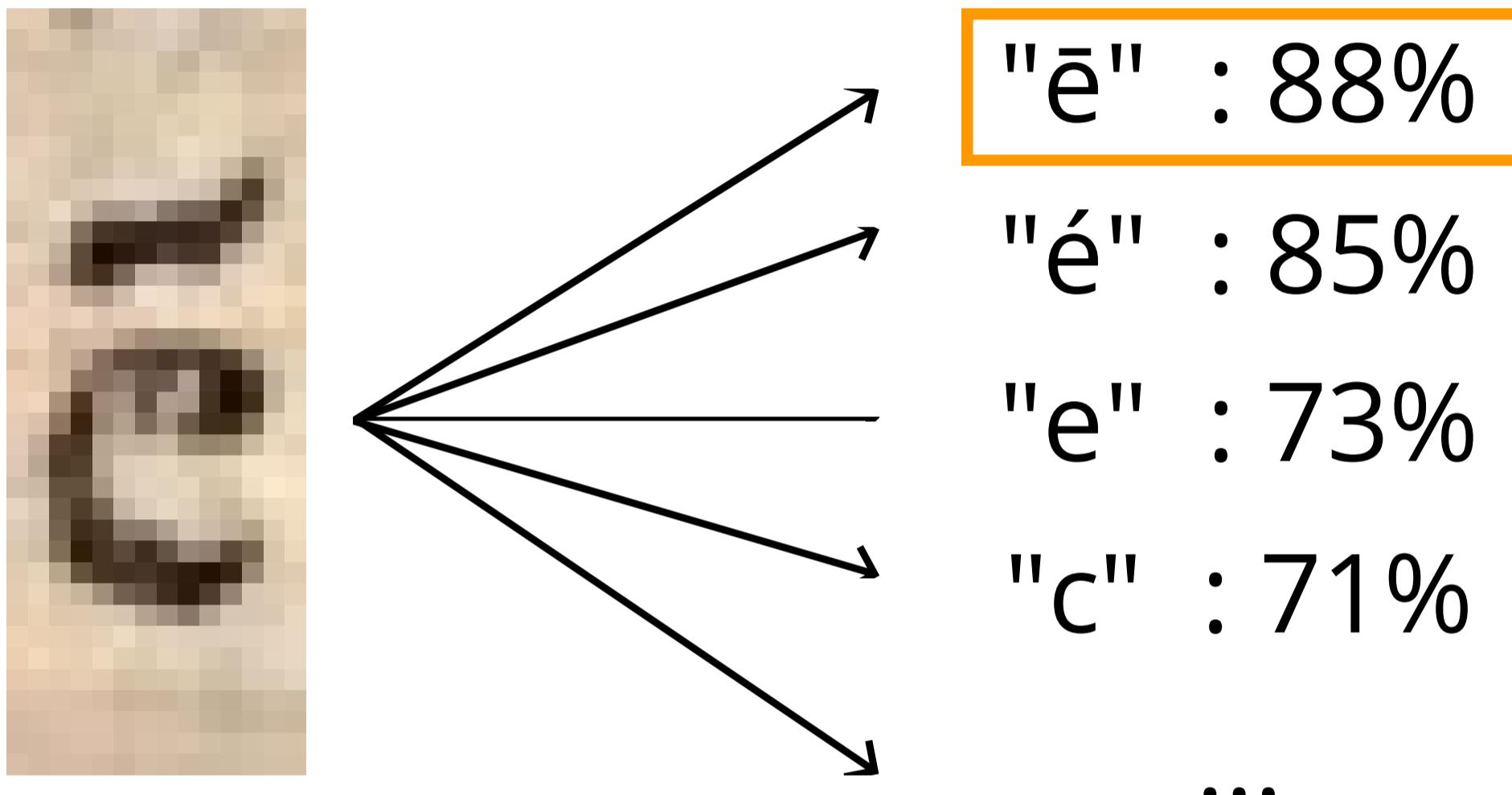
# Layout Analysis

BASF Aktiengesellschaft		
<b>BASF</b>	<b>BASF</b>	<b>BASF</b>
<b>Sitz:</b> 6700 Ludwigshafen (Rhein) <b>Telefon:</b> (06 21) 6 01 <b>Telex:</b> 4 64 811	<b>Sitz:</b> 6700 Ludwigshafen (Rhein) <b>Telefon:</b> (06 21) 6 01 <b>Telex:</b> 4 64 811	<b>Sitz:</b> 6700 Ludwigshafen (Rhein) <b>Telefon:</b> (06 21) 6 01 <b>Telex:</b> 4 64 811
<b>Vorstand:</b> Prof. Dr. rer. nat. Matthias Seestäder, Ludwigshafen (Rhein); Vors.; Dr. rer. nat. Hans Mödl, Ludwigshafen (Rhein); stellv. Vors.; Dr. rer. nat. Hans Albers, Ludwigshafen (Rhein); Dr. rer. pol. Ernst Deneel, Ludwigshafen (Rhein); Dr.-Ing. Erich Henkel, Ludwigshafen (Rhein); Dr. rer. nat. Wolfgang Jenisch, Ludwigshafen (Rhein); Prof. Dr.-Ing. Horst Pommier, Ludwigshafen (Rhein); Dr.-Ing. Karl August Weijen, Ludwigshafen (Rhein); Dr. rer. nat. Herbert Willemsen, Ludwigshafen (Rhein); Hans Joachim Witt, Ludwigshafen (Rhein). <b>Aufsichtsrat:</b> Prof. Dr. phil. nat. Bernhard Tinn, Heidelberg; Vors.; Werner Titz, Innsbruck, stellv. Vors. +; Dr. rer. iur. Wolfgang Arend, Ludwigshafen +;	<b>Vorstand:</b> Prof. Dr. rer. nat. Matthias Seestäder, Ludwigshafen (Rhein), Vors.; Dr. rer. nat. Hans Mödl, Ludwigshafen (Rhein); stellv. Vors.; Dr. rer. nat. Hans Albers, Ludwigshafen (Rhein); Dr. rer. pol. Ernst Deneel, Ludwigshafen (Rhein); Dr.-Ing. Erich Henkel, Ludwigshafen (Rhein); Dr. rer. nat. Wolfgang Jenisch, Ludwigshafen (Rhein); Prof. Dr.-Ing. Horst Pommier, Ludwigshafen (Rhein); Dr.-Ing. Karl August Weijen, Ludwigshafen (Rhein); Dr. rer. nat. Herbert Willemsen, Ludwigshafen (Rhein); Hans Joachim Witt, Ludwigshafen (Rhein). <b>Aufsichtsrat:</b> Prof. Dr. phil. nat. Bernhard Tinn, Heidelberg; Vors.; Werner Titz, Innsbruck, stellv. Vors. +; Dr. rer. iur. Wolfgang Arend, Ludwigshafen +;	<b>Vorstand:</b> Prof. Dr. rer. nat. Matthias Seestäder, Ludwigshafen (Rhein), Vors.; Dr. rer. nat. Hans Mödl, Ludwigshafen (Rhein); stellv. Vors.; Dr. rer. nat. Hans Albers, Ludwigshafen (Rhein); Dr. rer. pol. Ernst Deneel, Ludwigshafen (Rhein); Dr.-Ing. Erich Henkel, Ludwigshafen (Rhein); Dr. rer. nat. Wolfgang Jenisch, Ludwigshafen (Rhein); Prof. Dr.-Ing. Horst Pommier, Ludwigshafen (Rhein); Dr.-Ing. Karl August Weijen, Ludwigshafen (Rhein); Dr. rer. nat. Herbert Willemsen, Ludwigshafen (Rhein); Hans Joachim Witt, Ludwigshafen (Rhein). <b>Aufsichtsrat:</b> Prof. Dr. phil. nat. Bernhard Tinn, Heidelberg; Vors.; Werner Titz, Innsbruck, stellv. Vors. +; Dr. rer. iur. Wolfgang Arend, Ludwigshafen +;
<b>Dividenden auf Stammaktien:</b> 1973: 8 % (Div. Sch. Nr. 24) 1974 u. 1975: je 4 % (Div. Sch. Nr. 25, 26) 1976: 6 % (Div. Sch. Nr. 27) 1977: 6,- DM + 1,38 DM St.G., (Div. Sch. Nr. 28) <b>Aus den Gewinn- und Verlustrechnungen:</b> <b>Aus den Bilanzen (in 1.000 DM):</b> 1976 1977 Anlagevermögen 4 395 5 309 Umlaufvermögen 16 240 16 718 Übriges Vermög. 519 717 Eigenkapital 5 117 5 117 Grundkapital 3 815 3 815 Umsatzerlöse 17 918 22 106 Materialaufwand 7 316 8 319 Personalaufwand 9 297 10 389 Abschreibungen 618 767 BGFV-Steuern 299 329 Jahresüberschuss 241 184	<b>Dividenden auf Stammaktien:</b> 1973: 8 % (Div. Sch. Nr. 24) 1974 u. 1975: je 4 % (Div. Sch. Nr. 25, 26) 1976: 6 % (Div. Sch. Nr. 27) 1977: 6,- DM + 1,38 DM St.G., (Div. Sch. Nr. 28) <b>Aus den Gewinn- und Verlustrechnungen:</b> <b>Aus den Bilanzen (in 1.000 DM):</b> 1976 1977 Anlagevermögen 4 395 5 309 Umlaufvermögen 16 240 16 718 Übriges Vermög. 519 717 Eigenkapital 5 117 5 117 Grundkapital 3 815 3 815 Umsatzerlöse 17 918 22 106 Materialaufwand 7 316 8 319 Personalaufwand 9 297 10 389 Abschreibungen 618 767 BGFV-Steuern 299 329 Jahresüberschuss 241 184	<b>Dividenden auf Stammaktien:</b> 1973: 8 % (Div. Sch. Nr. 24) 1974 u. 1975: je 4 % (Div. Sch. Nr. 25, 26) 1976: 6 % (Div. Sch. Nr. 27) 1977: 6,- DM + 1,38 DM St.G., (Div. Sch. Nr. 28) <b>Aus den Gewinn- und Verlustrechnungen:</b> <b>Aus den Bilanzen (in 1.000 DM):</b> 1976 1977 Anlagevermögen 4 395 5 309 Umlaufvermögen 16 240 16 718 Übriges Vermög. 519 717 Eigenkapital 5 117 5 117 Grundkapital 3 815 3 815 Umsatzerlöse 17 918 22 106 Materialaufwand 7 316 8 319 Personalaufwand 9 297 10 389 Abschreibungen 618 767 BGFV-Steuern 299 329 Jahresüberschuss 241 184
<b>Dividenden auf Stammaktien:</b> 1973: 8 % (Div. Sch. Nr. 24) 1974 u. 1975: je 4 % (Div. Sch. Nr. 25, 26) 1976: 6 % (Div. Sch. Nr. 27) 1977: 6,- DM + 1,38 DM St.G., (Div. Sch. Nr. 28)	<b>Dividenden auf Stammaktien:</b> 1973: 8 % (Div. Sch. Nr. 24) 1974 u. 1975: je 4 % (Div. Sch. Nr. 25, 26) 1976: 6 % (Div. Sch. Nr. 27) 1977: 6,- DM + 1,38 DM St.G., (Div. Sch. Nr. 28)	<b>Dividenden auf Stammaktien:</b> 1973: 8 % (Div. Sch. Nr. 24) 1974 u. 1975: je 4 % (Div. Sch. Nr. 25, 26) 1976: 6 % (Div. Sch. Nr. 27) 1977: 6,- DM + 1,38 DM St.G., (Div. Sch. Nr. 28)

- text vs. image classification
- header, footer, headings
- multi-columns, reading order
- line recognition

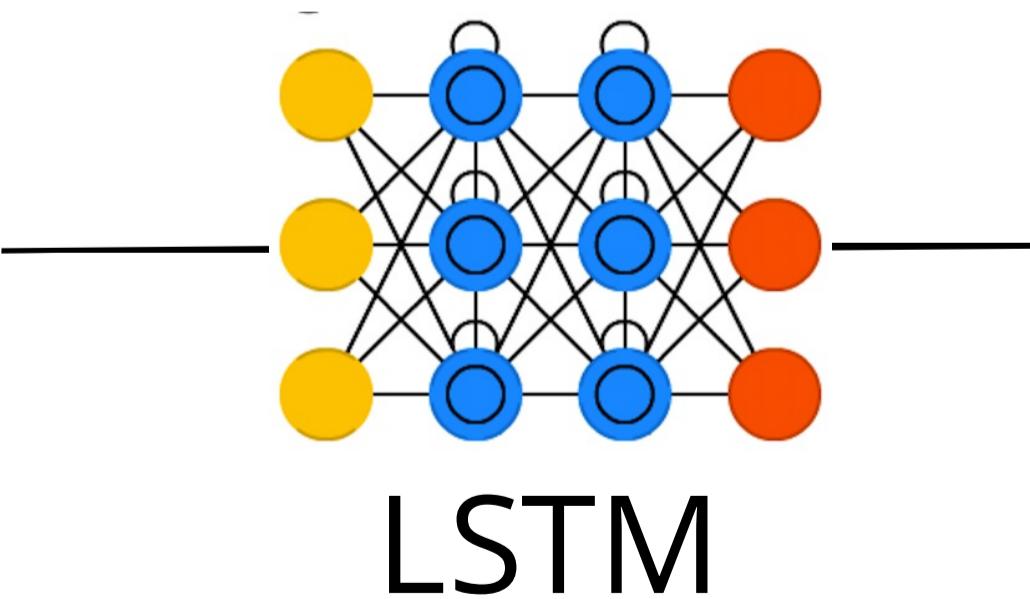
# Text Recognition

## a) character-based recognition



## b) line-based recognition

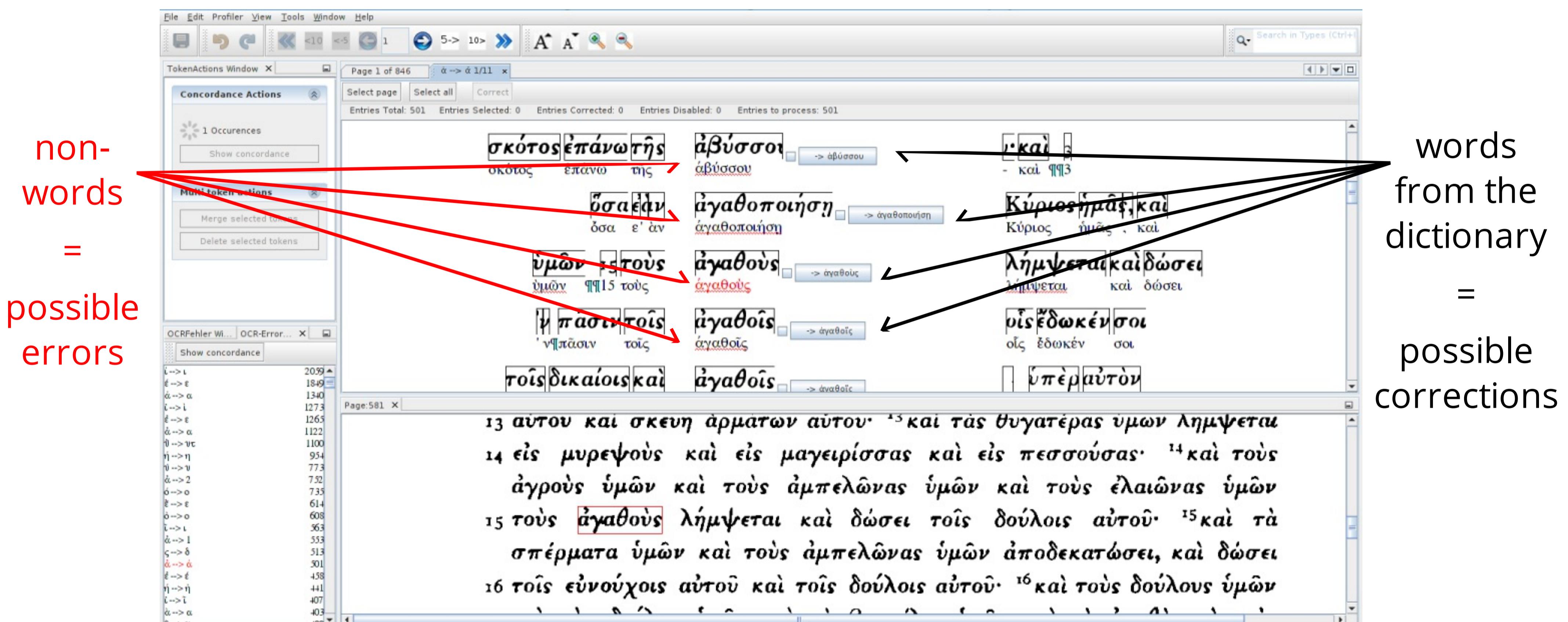
mit Weglassung solcher Verse



"mit Weglassung  
solcher Verse"

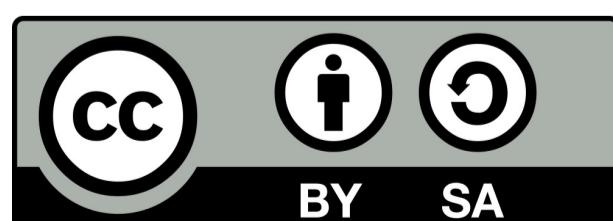
# Computerlinguistical Methods

- dictionary
- bigram, -trigrams, etc. for letters and words



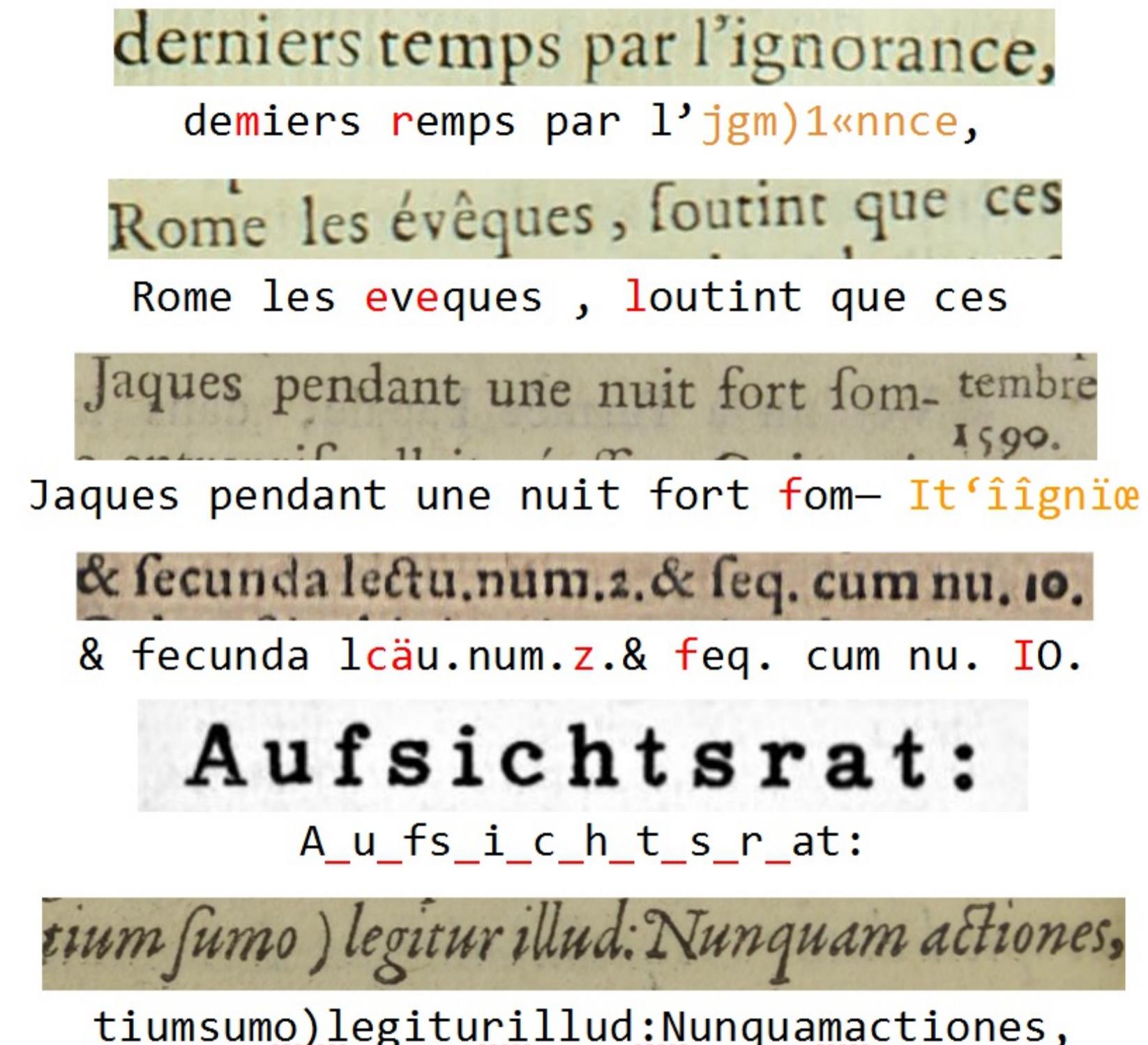
Screenshot from PoCoTo used as CC-BY-SA published in:

CIS München (2016): [Abschlussbericht zum Projekt "Ausbau und Erweiterung eines Open-Source-Tools zur Nachkorrektur historischer OCR-erfasster Texte"](#) der CLARIN-D Facharbeitsgruppe 4-3 "Klassische Philologie"



# Recognition Errors

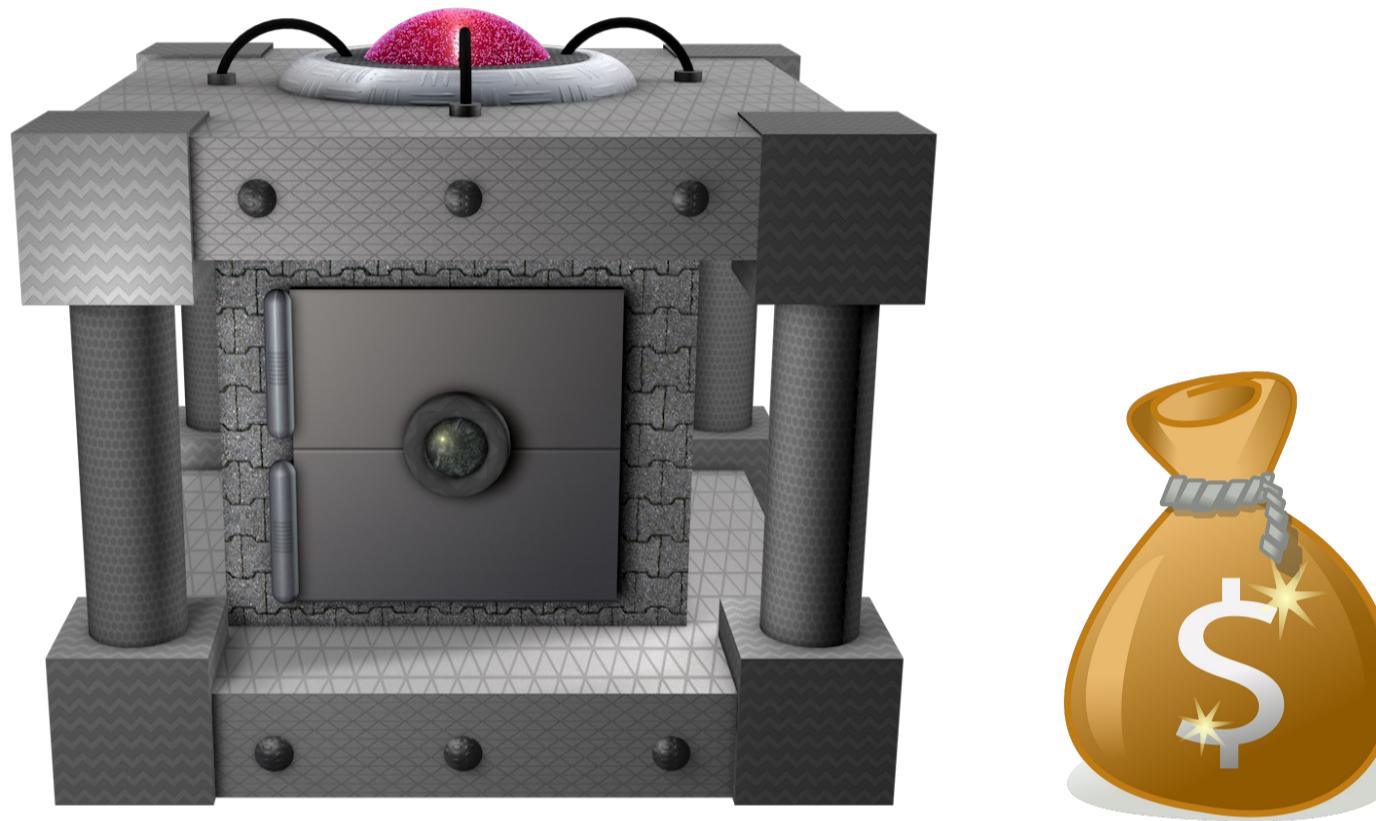
- OCR results have errors
- errors can occur in each step
  - scanning errors
  - segmentation/layout errors
  - recognition errors
  - errors in dictionaries
  - untrained characters



*Advise:* Judge the errors with regard to your application  
(fuzzy search, topic modeling, extracting exact numbers)

# **OCR-Software**

# Commercial OCR Software



## ABBYY Finereader

e.g. FineReader Engine 11 CLI for Linux  
(on one server/pc): 120'000 pages / year  
for 999 EUR

# Open Source OCR Software



## Tesseract

- started 1985 by HP Labs
- since 2006 Open Source
- supported by Google

## Ocroptus

- started 2007
- founded and maintained by Prof. Breuel (DFKI, Google, Nvidia)

etc.

# ABBYY Finereader

- Normally good results
- Closed source, limited options to change behaviour
- Strong emphasize on language-dependent dictionaries

```
abbyyocr11 -rl German \
-if input.jpg \
-f PDF -of output.pdf
```

# Tesseract

- Until 2016 character-based text recognition only, now also neural-network-based text recognition
- Less emphasize on language-dependent dictionaries
- [github.com/tesseract-ocr/tesseract](https://github.com/tesseract-ocr/tesseract), part of linux distrib.
- For Windows: [github.com/UB-Mannheim/tesseract/wiki](https://github.com/UB-Mannheim/tesseract/wiki)
- For R: [github.com/ropensci/tesseract](https://github.com/ropensci/tesseract)

```
tesseract input.jpg output \
-l eng+deu \
--oem 1 --psm 7 \
hocr
```

# OCRopus

- neural network algorithm since 2013
- training is key feature
- different models for scripts (not languages)
- no dictionary
- modular scripts (Unix philosophy)

```
./ocropus-nlbin tests/ersch.png  
./ocropus-gpageseg ersch/*.bin.png  
./ocropus-rpred ersch/*/*.bin.png \  
-m models/fraktur.pyrnn.gz  
./ocropus-hocr ersch/*.bin.png
```

# OCR Fileformats

- recognized text
- position of the words, lines, characters (bounding boxes)
- confidence values
- text direction, recognized language, formats, ...

e.g. hocr file:

```
<p class='ocr_par' lang='deu' title="bbox930">
...
<span class='ocr_line' title="bbox 348 797 1482 838; baseline -0.009 -6">
    <span class='ocrx_word' title='bbox 348 805 402 832; x_wconf 93'>Die</span>
    <span class='ocrx_word' title='bbox 421 804 697 832; x_wconf 90'>Darlehenssumme</span>
    <span class='ocrx_word' title='bbox 717 803 755 831; x_wconf 96'>ist</span>
    <span class='ocrx_word' title='bbox 773 803 802 831; x_wconf 96'>in</span>
    <span class='ocrx_word' title='bbox 821 803 917 830; x_wconf 96'>ihrem</span>
    <span class='ocrx_word' title='bbox 935 799 1180 838; x_wconf 95'>ursprünglichen</span>
    <span class='ocrx_word' title='bbox 1199 797 1343 832; x_wconf 95'>Umfänge</span>
    <span class='ocrx_word' title='bbox 1362 805 1399 823; x_wconf 95'>zur</span>
    <span class='ocrx_word' title='bbox 1417 x_wconf 96'>ver-</span>
</span>
...

```

Other OCR-formats: ALTO, Page XML, ABBYY XML, TEI, GCV

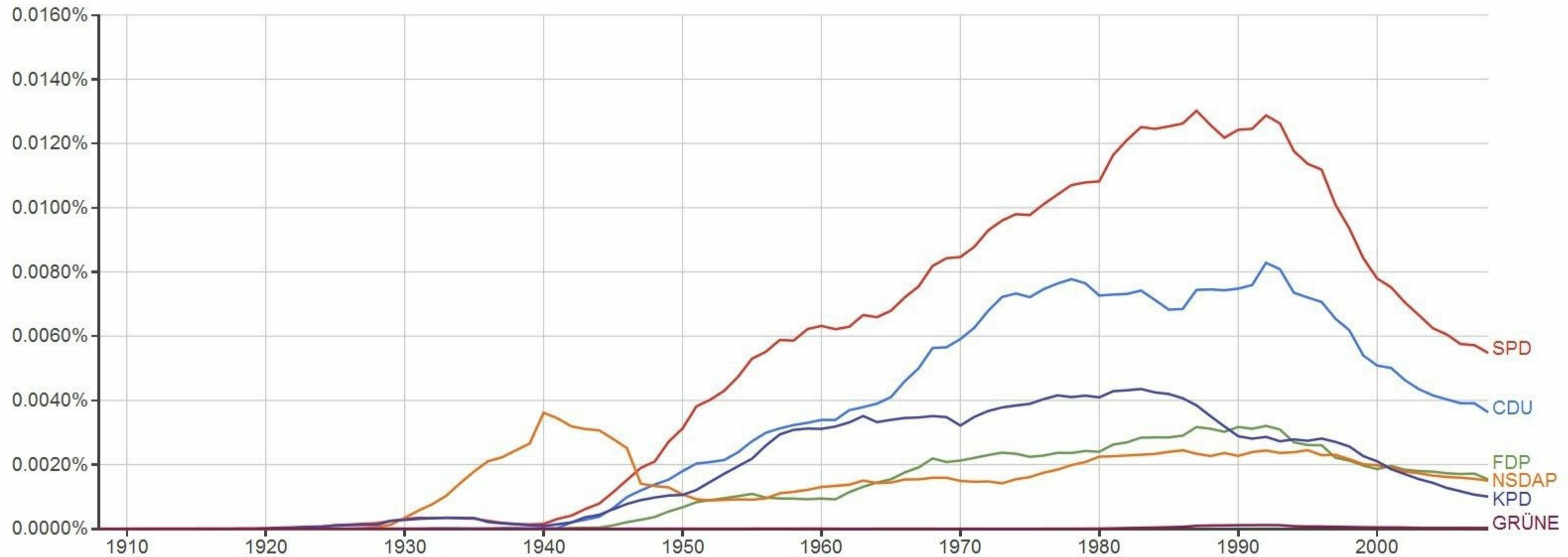
<https://github.com/UB-Mannheim/ocr-fileformat>

# Applications

# Ngram Viewer (Google Books)

## Google Books Ngram Viewer

Graph these comma-separated phrases: CDU,SPD,FDP,NSDAP,KPD,GRÜNE  case-insensitive  
between 1908 and 2008 from the corpus German with smoothing of 3



-> [View this query online](#)

# Number of Females in the Supervisory Board of DAX-30 companies 1979-1999

1. Go to the "Aktienführer Datenarchiv" and there to "Export"
2. Increase number of results to 50, search for "DAX", click on select all visible (38 results)
3. Adjust the year range
4. Select the category "Supervisory Board"
5. Export the CSV data
6. Open in Excel, mark the female names
7. Finally make a pivot table

The screenshot shows a list of 38 selected companies from a total of 1186. The companies listed are: SAP Aktiengesellschaft Systeme, Anwendungen, Produkte in der Datenverarbeitung (DAX 18.09.1995-), GEA Aktiengesellschaft (DAX 03.09.1990 - 18.11.1996), adidas Aktiengesellschaft (DAX 22.06.1998-), Metro Aktiengesellschaft (DAX 22.07.1996 - 24.09.2012), and Deutsche Telekom AG (DAX 18.11.1996-). Below the list are buttons for 'select visible' and 'clear selection', and a note stating 'Sie haben 38 von 1186 Unternehmen ausgewählt'. Navigation buttons for 'Previous', '1', and 'Next' are also present.

The screenshot shows the 'Jahre' (Years) section with 'Erstes Jahr' set to 1979 and 'Letztes Jahr' set to 1999. Below it is the 'Datenkategorie wählen' (Select Data Category) section, which has 'Aufsichtsrat / Supervisory Board' selected. There are options for separator selection ('Felder getrennt durch') including Komma (,), Semikolon (;), Verkettungszeichen (), Tab, and UTF8-BOM explizit schreiben (z.B. für Excel). A 'UTF8-BOM explizit schreiben' checkbox is checked. At the bottom is a 'Exportieren' (Export) button.

Unternehm	Profil-ID	Unternehmensname	Erscheinun	Vorname (i_w)	Titel (Aufsi	Ort (Aufsic	Funktion (F	Bemerkung	Rang (Aufs
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Franz Heinrich	m	Düsseldorf	Vors.		1
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Konrad	m	München	stellv. Vors., Arbeitnehmervertreter		2
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Hermann	m	Hamburg	Arbeitnehmervertreter		3
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	J. R. M.	m	Prof. Dr.	Amsterdam		4
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Hans	m	Dr.	Frankfurt		5
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Friedrich Karl	m	Dr.	Düsseldorf		6
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Jörg A.	m		Duisburg		7
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Erich	m		Pforzheim	Arbeitnehmervertreter	8
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Karl	m	Dr.	Hamburg		9
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Axel	m	Frankfurt	Arbeitnehmervertreter		10
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Hans L.	m	Stuttgart			11
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Karl	m	Düsseldorf	Arbeitnehmervertreter		12
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Josef	m	Köln	Arbeitnehmervertreter		13
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Irene	w	Salzgitter	Arbeitnehmervertreter		14
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Marion	w	Essen	Arbeitnehmervertreter		15
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Peter	m	Dr.rer.pol.	München		16
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Heinz	m	Düsseldorf	Arbeitnehmervertreter		17
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Günter	m	Düsseldorf	Dipl.-Kfm.		18
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Lothar	m	Köln	Arbeitnehmervertreter		19
1141	19.790.269	Deutsche Bank Aktiengesellschaft	1979	Hannelore	w	Düsseldorf			20
1141	19.800.172	Deutsche Bank Aktiengesellschaft	1980	Franz Heinrich	m	Düsseldorf	Vors.		1
1141	19.800.172	Deutsche Bank Aktiengesellschaft	1980	Konrad	m	München	stellv. Vors., Arbeitnehmervertreter		2

# Number and age of German voters for EU vote 1989

1. Go to [digizeitschriften.de](http://digizeitschriften.de) and then to the [Statistisches Jahrbuch für die Bundesrepublik Deutschland 1990](#)
  2. Download the pdf of the chapter "Wahlen" starting from page 76
  3. Open the pdf in the PDF X Change Viewer, run OCR and save it (or the alternatives you heard before)
  4. Download **Tabula** <http://tabula.technology/>,  
install it and run it
  5. Open pdf in Tabula, select table and extract data as csv
- (\*) The quality is here not yet optimal, but it shows the possibilities of the tools and data around OCR.



# Number of German Emmigrants from 1870 until 1880

1. Go to the Reichsanzeiger
2. Search for "Auswanderer"
3. Be lucky
4. Go to the result

## Search in Reichsanzeiger fulltext

Search text: Auswanderer

Max. number of errors:

0  1  2

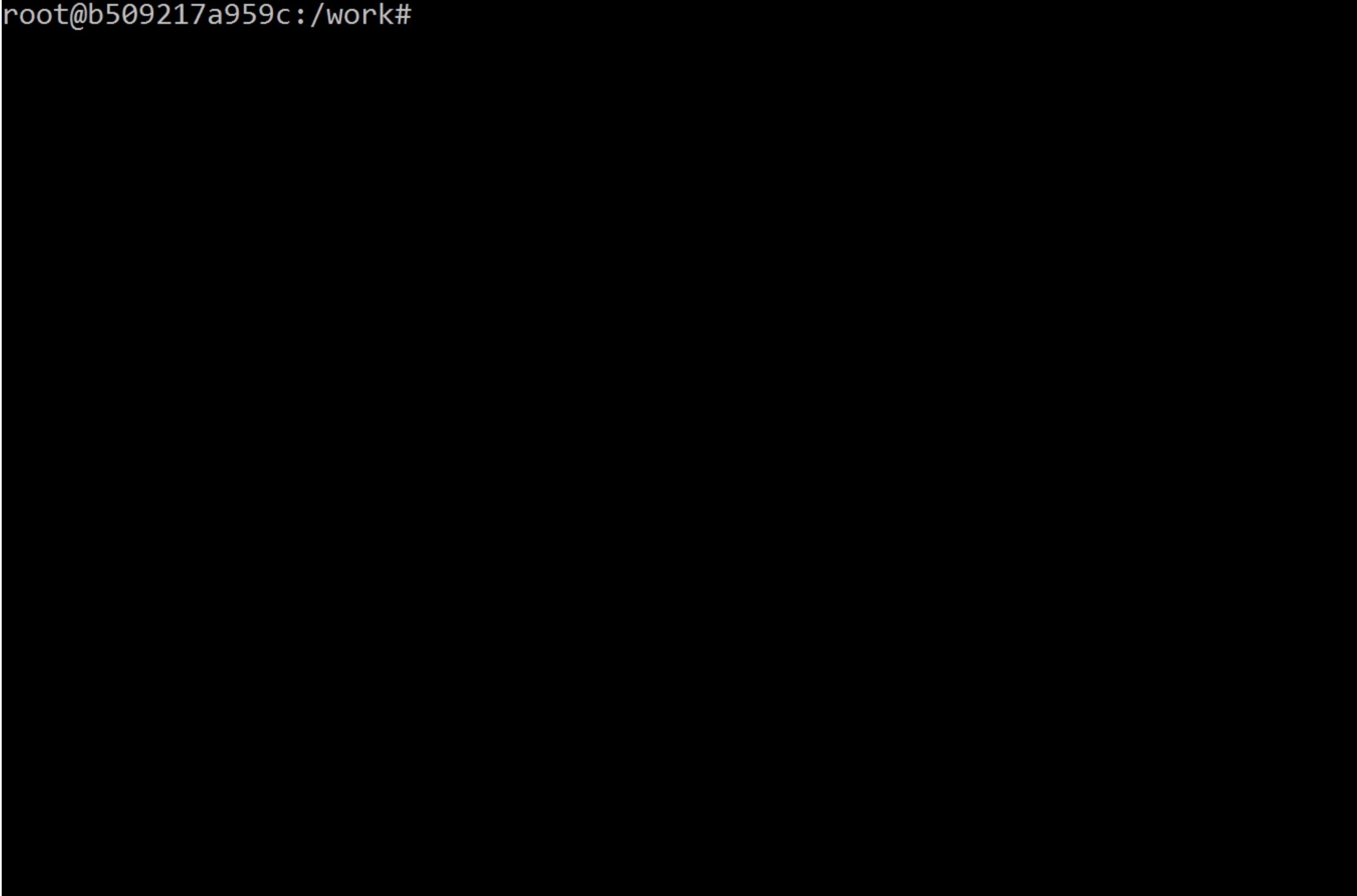
Submit Query

[duplicat/025-1903/1145.txt: Auswanderern sowie über den Entzug von Vorschriften über](#)  
[film/001-9882/0162.txt: \(93\); solchen Arbeiten könnten die Auswanderer gro-](#)  
[film/001-9882/0211.txt: sellschaft zur Unterstützung teutscher Auswanderer nach](#)  
[film/002-9883/0318.txt: um mehr den armen Auswanderern theils die s-U](#)  
[film/002-9883/0318.txt: der Auswanderer mit Enthusiasmus](#)  
[film/003-1881/0025.txt: liebe, 42 413 meibl\\_icbe, zusammen 106191 Auswanderer an; und zwar](#)  
[film/003-1881/0025.txt: Das Steigen und Sinken der Auswandererzahlen bringt wesent-](#)  
[film/003-1881/0025.txt: Vereinigten Staaten ziehen immer noch alle unsere Auswanderer bis](#)  
[film/003-1881/0032.txt: Auswanderung derselben nach Amerika unausführbar.](#)  
[film/004-1882/0006.txt: unter den auswanderern sind 106191 auswanderer aus dem](#)

1871 . . .	75 912	1876 . . .	28 368
1872 . . .	125 650	1877 . . .	21 964
1873 . . .	103 638	1878 . . .	24 217
1874 . . .	45 112	1879 . . .	33 327
1875 . . .	30 773	1880 . . .	106 191.
Das Steigen und Sinken der Auswandererzahlen hängt wesentlich von den günstigen oder ungünstigen Berichten ab, welche über die wirtschaftlichen Verhältnisse in den Vereinigten Staaten und die Aussicht auf gutes Fortkommen dort zu uns gelangen, denn die Vereinigten Staaten ziehen immer noch alle unsere Auswanderer bis auf einen kleinen Bruchteil an sich. Von den nachgewiesenen 106 191 Auswanderern des Jahres 1880 gingen nach:			
den Vereinigten Staaten . . .	103 116		
Brasilien . . . . .	2 119		
anderen amerikanischen Staaten . . . . .	761		
Australien . . . . .	132		
Asien . . . . .	36		
Afrika . . . . .	27		

# Discussion, Questions?

```
root@b509217a959c:/work#
```



OCRopus run-test executes nlbin, gpageseg, rpred

# List of Images

- Slide 1: <https://pixabay.com/de/hong-kong-stadt-st%C3%A4dtischen-1990268/> (CC0)
- Slide 3.2: Copyright User (2013-06): Text and Data Mining (Original Illustration by Davide Bonazzi)  
<http://copyrightuser.org/topics/text-and-data-mining/> (CC-BY)
- Slide 3.4: Copyright User (2013-06): Text and Data Mining (Original Illustration by Davide Bonazzi)  
<http://copyrightuser.org/topics/text-and-data-mining/> (CC-BY), <https://pixabay.com/de/b%C3%BCcher-stapelbildung-lesung-41930/> (CC0), <https://pixabay.com/de/zeitung-artikel-zeitschrift-154444/> (CC0)
- Slide 4.3: The two images of our scanners are made by the Mannheim University Library 2017 (can be used as CC-BY)
- Slide 5.2: Baierer, Konstantin; Zumstein, Philipp (2016). Verbesserung der OCR in digitalen Sammlungen von Bibliotheken. *027.7 Zeitschrift für Bibliothekskultur / Journal for Library Culture*, v. 4, n. 2, p. 72-83.  
<https://doi.org/10.12685/027.7-4-2-155> (CC-BY)
- Slide 5.3 and 5.4: Images created for this talk (CC0)
- Slide 5.5: LSTM <http://www.asimovinstitute.org/neural-network-zoo/> (CC0)
- Slide 5.6: Screenshot from PoCoTo (CC-BY-SA) published in: CIS München (2016): [Abschlussbericht zum Projekt "Ausbau und Erweiterung eines Open-Source-Tools zur Nachkorrektur historischer OCR-erfasster Texte"](#) der CLARIN-D Facharbeitsgruppe 4-3 "Klassische Philologie"
- Slide 5.7: Baierer, Konstantin; Zumstein, Philipp (2016). Verbesserung der OCR in digitalen Sammlungen von Bibliotheken. *027.7 Zeitschrift für Bibliothekskultur / Journal for Library Culture*, v. 4, n. 2, p. 72-83.  
<https://doi.org/10.12685/027.7-4-2-155> (CC-BY)
- Slide 6.2: <https://pixabay.com/de/beutel-geld-reichtum-einnahmen-147782/> (CC0),  
<https://pixabay.com/de/quell-offene-software-offene-software-1518247/> (CC0),  
<https://pixabay.com/de/sicher-metall-metallischen-ger%C3%A4t-298244/> (CC0)
- Several logos and screenshots